学校编码：10384　　分类号_____密级 _____

学号：14220051300748　　UDC _____

厦门大学

硕 士 学 位 论 文

数据挖掘中金融时间序列的粗糙聚类分析

Rough Clustering of Financial Time Series in Data Mining

吴晓彬

指导教师姓名：　朱 建 平 教授

专 业 名 称：　统　　计　　学

论文提交日期：　2008 年　　月

论文答辩日期：　2008 年　　月

学位授予日期：　2008 年　　月

答辩委员会主席：_____

评　　阅　　人：_____

2008 年　　月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（ ），在 年解密后适用本授权书。

2、不保密（ ）

（请在以上相应括号内打“ √ ”）

作者签名：　　日期：　　年 月 日

导师签名：　　日期：　　年 月 日

内容摘要

传统统计分析与现代金融计量经济方法研究时间序列的主要思路是建立基于严格数学推导下的统计模型并对其进行参数估计与数据检验，目前已建立起一套较为成熟的理论体系。但该方法既依赖于苛刻的假设条件，又要求所有数据都符合一个固定的数学模型，显得过于牵强。数据挖掘研究时间序列的思路则不同，它由数据直接驱动建立模型，克服了上述的缺陷。

时间序列数据挖掘已是当前的研究热点之一，人们也取得不少的研究成果，但对于时间序列相似性度量这一关键难题一直未能得到较好的解决，而很多时序挖掘方法都是建立在相似性的基础上，显然时间序列相似性度量直接影响着这些时序挖掘方法的结果，为此本文首先就该关键的基础性问题展开研究，进一步讨论了该度量方法在序列挖掘中的应用。由于数据挖掘方法众多，本文不可能一一涉及，所以只针对聚类分析进行深入的探讨。聚类分析不仅是数据挖掘的重要组成部分，同时也是多元统计分析的重要方法，在实际中有广泛的运用。本文绕开了已有较多成熟方法的硬聚类，而深入地研究了一种软聚类——粗糙聚类的方法及其在时间序列挖掘中的应用，同时从侧面反映了本文度量序列相似性方法的实用性。全文的主要工作及创新可归纳为以下几点。

首先，结合小波分析的思想方法，提出一种基于小波多尺度变换的时间序列相似性度量方法，并通过金融时间序列的实例研究，说明该方法全面考虑了影响序列相似性度量的各种因素，很好地克服了已往方法无法兼顾序列整体形状轮廓与细节差异的缺陷。

其次，在相似性度量方法的基础上，研究了序列粗糙聚类方法，通过金融实证研究表明粗糙聚类方法的优点。并深入研究了以下三个问题：（1）建立粗糙聚类质量指标，并研究不同阈值参数对聚类结果的影响；（2）将粗糙聚类法与层次聚类法进行整合，各取所长；（3）将软聚类转化为硬聚类，通过迭代剔除法对粗糙聚类结果精简化，并与之前聚类结果进行比较，说明其可行性。

最后，本文模型方法尚无现成的软件模块实现，故本文还给出 Matlab 软件上具体实现的参考程序，结合实证研究取得较好的效果。

关键词：数据挖掘；时间序列；相似性度量；小波分析；粗糙聚类；

Abstract

Based on strict mathematical conduction and then to conduct parameters estimation and inference, traditional statistics and modern financial econometrics, in which theory frameworks have been built up for years, are to establish statistical models. However, such methods seem unfit due to its dependence on strict hypothesis and importuning all data of series to meet modeling requirements. Data mining techniques overcome this kind of shortage in a way of establishing models motivated by data.

Time series data mining is popular today, and many achievements have been made. Whereas, appropriate solution of measuring similarity still lacks of attention, which lays the foundation of several methods in series mining. Apparently, similarity measurement in time series does affect mining results. This dissertation aims at such pivotal issue as well as its applications in series mining, particularly, clustering analysis. Instead of hard clustering, this dissertation introduces a soft clustering method——Rough Clustering method, which can reflect the practicability of the new method on measuring similarity of time series. Main works and innovations of this dissertation are summarized as:

Firstly, a method to measure similarity of time series based on multi-scale wavelet transformation is presented with the idea of wavelets analysis. And financial time series cases study is also conducted to show that this method considers all the factors affecting the measuring similarity of series and effectively overcomes the shortage of existent methods that fail to balance between outline and detail differences of series.

Secondly, discusses rough clustering of sequences and shows its advantages through financial cases study. Furthermore, analysis on three issues as follow is considered: (1) to discuss the impact of threshold parameters on clustering results by establishing the quality indicators for rough clustering; (2) to integrate the rough clustering and hierarchical clustering so that we can make most of their advantages; (3) to transfer soft clustering into hard clustering, to condense the results of rough

clustering by the iteratively-removed-method, and to show its feasibility by comparing with original results.

Finally, we also discuss the algorithms used in these methods, and share programming code in form of Matlab. Results from empirical research are convincible.

Key Words: Data Mining; Time Series; Similarity Measurement; Wavelet Analysis; Rough Clustering

目 录

Contents

第1章 绪论

1.1 数据挖掘的兴起

1.1.1 数据挖掘的重要性

数据挖掘是信息领域发展最快的技术之一，很多不同领域的专家，如统计学家、数据库专家等，都从中获得了发展的空间。随着计算机技术，特别是数据库技术的快速发展和广泛应用，各行各业积累的数据日益膨胀，数据量达到 GB 甚至 TB 级，传统的数据处理方式已很难充分利用蕴藏在这些数据中的有用知识，从而导致了“数据丰富，但信息贫乏”的现象[1]，激增的数据唤起了人们对挖掘其中所隐藏知识的需求，于是数据挖掘这一整合多种分析手段，从大量数据中发现有用知识的方法就应运而生，并在使用中得以蓬勃发展。

1995 年在加拿大蒙特利尔召开的第一届知识发现和数据挖掘国际会议上，“数据挖掘”概念第一次由 Usama Fayaad 提出，这次会议一直被认为是该领域的主要会议之一。数据挖掘“是一门能对观测到的数据集(经常是很庞大的)进行分析，目的是发现未知的关系和以数据拥有者可以理解并对其有价值的新颖方式来总结数据的技术”[2]。或者说“是从数据集中识别有效的、新颖的、潜在有用的，以及最终可理解的模式的高级处理过程。”[3]。总的说来，数据挖掘是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但是又潜在有用的信息和知识的过程，它包括数据清理、集成、选择、变换、挖掘、模式评估、知识表达等过程。它应用各种方法从数据序列中发现隐含的规律和模式，这些方法可能来自于各个领域，比如统计学、人工智能、神经网络、粗糙集、支持向量机、模糊逻辑等等，甚至也包括其它新鲜的方法。其功能主要包括概念描述和可视化、关联分析、分类和预测、异常分析、趋势分析等[4]。

如今数据挖掘技术已在购物篮分析、客户关系管理、产品质量分析、基因工程研究、Internet 站点访问模式发现等许多领域得到成功应用。根据 Gartner Group 的一次高级技术调查，其报告将数据挖掘和人工智能列为“将对工业产生深远影响的五大关键技术”之首，并且还将并行处理体系和数据挖掘列为未来五年内投资焦点的十大新兴技术前两位[9]。

1.1.2 统计学与数据挖掘的相互影响

数据挖掘中有很多思想方法源自统计学，常见的数据挖掘软件都有提供统计分析功能，这对于数据挖掘的前期数据探索和数据挖掘之后对数据进行总结和分析都是必不可少的。统计分析中诸如时间序列分析、假设检验、相关性分析、方差分析、线性预测等方法都有助于数据挖掘前期对数据进行探索，发现挖掘的主题，定位挖掘的目标，确定挖掘涉及的变量，对数据源进行抽样等等。所有这些前期的探索工作都对数据挖掘的效果与质量产生重大影响，且数据挖掘的结果也需要统计分析的描述功能进行具体描述，使数据挖掘的结果能被用户所了解。

数据挖掘并不是为了替代传统的统计分析技术，而是统计分析方法的拓展与延伸，二者相辅相成，相互促进。统计分析中的许多技术都是建立在完善的数学理论和高超的建模技巧基础上，其分析与预测的准确程度还是令人满意的，但对于使用者的知识要求比较高。而随着计算机能力的不断发展，数据挖掘可以利用相对简单和固定程序完成同样的任务。新兴的计算方法如决策树、神经网络使人们不需了解到其内部复杂的原理也能通过这些方法获得良好的分析和预测效果。数据挖掘作为多门学科的综合，其分析问题的思维不再局限于传统统计分析的模式，它已经从机器学习那里继承了实验的态度，这与传统统计分析的本质区别是数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识。数据挖掘得到的信息应具有启发性，有效性和实用性三个特征。由此可见，数据挖掘方法和统计模型分析方法尽管其目的和出发点相似，但由于解决问题的思路存在本质上的区别，因而两类方法对数据规律的提取形式和效果都有所不同，因而它们是两类不同、可相互弥补不足的分析方法[9]。对两者开展深入研究，都具有非常重要的意义，至于二者谁是谁的分支这类的争论，已经没有多大意义，实际上，由于挖掘方法和统计方法间的联系，特别是数据挖掘极富包容性的开放式思维风格，它经常借鉴和引用统计方法的很多成果，而同时，数据挖掘的进一步发展也对统计学提出了更高的要求与挑战。

1.2 选题背景及意义

1.2.1 金融时间序列研究的重要性

时间序列是指按时间顺序排列的一组数据，是最常见的数据形式之一。在金

融、工业、医药、气象、计算机网络等十分广泛的领域，存在大量带有时间属性的数据。在数据挖掘领域内，对时间序列的关注也越来越多，针对时间序列的数据挖掘已成为一个新的热点。人们对时间序列的研究已经开展了很长时间，特别在利用时间序列数据来进行建模、预测等方面，已经取得了相当多的研究成果。而随着人工智能、机器学习等信息科学技术的飞速发展，对时间序列进行信息知识方面的研究也越来越广泛，比如时间序列的查询、编码、分类等等。因此时间序列数据挖掘自 20 世纪 90 年代以来得到了快速的发展。

在金融领域，数据绝大多数表现为时间序列数据，常见的如股票或期货市场中的各种价格、成交量、持仓量、收益率，货币市场中的利率，外汇市场中的汇率等数据都是以时间序列的形式记录保存的。本文选用金融时间序列数据作为挖掘对象，主要出于两方面的考虑：一是金融数据的质量高，而保证数据质量是进行数据挖掘的重要前提，所挖掘出的知识才有价值，否则将会是“垃圾进，垃圾出”的尴尬局面；二是金融在国民经济发展中占有举足轻重的地位，金融市场是国家乃至世界经济运行的核心，探析金融市场的变化规律、进行有效的金融管理、提高金融投资效率是各国政府与投资机构孜孜以求的目标之一。特别是在我国经济与世界经济相互融合程度越来越高，金融业面临着更大的、新的发展机遇和挑战的今天，以及在全球金融创新活动日新月异、各国金融市场联动效应不断增强的国际背景下，对金融市场本质规律的认识和把握更直接关系到金融市场的稳定、效率与安全。我国金融业经过多年来的高速发展，已经形成了相当规模的市场，金融机构的许多业务活动(如价格预测、客户分析、投资决策、风险管理等)都越来越依赖于对大量历史数据的分析，我国的投资者与金融机构也越来越清楚地认识到分析金融数据、从中挖掘出有价值的信息是其实现科学化管理决策的重要手段与“基础核心”工作。金融市场是一个非常庞大的系统，受多种因素影响，其运动规律极其复杂，而时间序列数据则是其综合外在表现形式。“本质决定现象，现象反映本质”，因此时间序列中必定蕴含了金融系统许多客观规律信息。从中挖掘出各种信息，更好地认识、掌握、并利用其规律无疑对金融投融资决策与风险管理活动具有特别重要的意义。

兰秋军等人分析指出，对金融数据的分析方法，主要分模型法和挖掘法两大类[9,10]。模型法是指在各种假设基础之上，建立数学模型，然后运用历史或当前

数据来进行决策与预测分析，最后根据有关金融理论得出结论。它是现代金融计量经济学理论中的重要内容，主要以数理统计模型为基础，实质就是研究如何构造一个与现实情况符合的预测模型，最大程度地减少预测误差。如美国经济学家 Engle 就因其在 1982 年对金融时间序列所提出的 ARCH 模型而荣获 2003 年度诺贝尔经济学奖。为了构建模型，许多假设条件是必须的。比如常用的 ARMA 模型要求时间序列是平稳的，并要求 ARMA 模型所产生的时间序列与观察序列间的误差相互独立，且呈正态分布。即便对目前研究较多的 ARCH、GARCH 类模型同样也脱离不了类似假设。在这些假设基础上建立起来的这些模型固然具有一种无与伦比的“简洁美”，然而，这些假设条件对许多实际情况来说却是非常“苛刻”的。而且，这种统计模型分析技术存在一个致命的缺陷，即它们总是着眼于所考察数据的全体。或者说，模型在构建过程中是以对“所有”考察数据的最佳适应为准则的。模型一旦构建出来，它将“适用”于数据序列的各个部分。显然，这里的一个隐含假定是金融时间序列是保持某种结构不变的。但现实金融市场是一个复杂系统，往往是以多种方式对外界作用起反应的。因而，金融时间序列的随机性是非常强的，以一个全局统计模型来囊括其运动规律太理想化，自然会造成较大的估计偏差，甚至得出一些错误的结论。正因为如此，依赖这种方法建立起来的许多模型在现实中往往失效，并可能带来无法估量的损失。一个令人瞩目的例子是，由美国著名经济学家、诺贝尔经济学奖获得者、期权定价理论的创立者——莫顿和舒尔斯管理的长期资本管理基金，在 1998 年的一次投资活动中惨遭失败，损失达 15～20 亿美元之巨，而不得不面临倒闭的结局。

挖掘法总的说来基于归纳推理的思维。“知识”之所以被发现是因为有足够多的数据支持它。因此它缺乏严格的理论支撑，而更多的是“经验”基础。更何况由于各种干扰因素的影响，数据中确实会存在“假”知识。因此它发现的“知识”一般还需通过其它手段进行验证。不过，它能发现“新”知识这一点非常重要，尽管发现出来的有些模式或规则目前也许难以理解，但它可能具有极好的“启发”价值。由于挖掘方法基于归纳的思想，它是直接以数据驱动的，因而它常常可以撇开一些假设条件，如不须正态假设、平稳假设、线性假设等等。挖掘工具开发好后，即使对挖掘理论不了解，用户也可以自助地对其感兴趣的数据进行挖掘，因而也容易使用，事实上数据挖掘很大程度上就是为这种便利性而提出来的。

对金融时间序列的挖掘不仅是有益的尝试，更是对金融计量学分析的良好补充，这也是本文的初衷。

1.2.2 小波分析的实用性

传统概率统计学下的时间序列分析，经过数十年的研究已经形成了自己的理论体系，但传统的方法多单独集中于时域或频域，而金融时间序列十分复杂，从单方面的时域或频域很难充分反映其特征，或者反映速度较慢，或者没有定位作用，因而分析金融时间序列应采用时频相结合的分析方法，这其中最引人注目的就是小波分析理论。

小波理论是目前国内外学术界高度关注的前沿领域，包含了极其丰富的数学内容，具有广泛使用的潜力，正在科学技术界掀起一场革命。在数学领域，它是泛函分析，Fourier 变换，样条分析，调和分析，数值分析的完美结合。在信号处理、图像处理、语音识别、模式识别、数据压缩、故障诊断、量子物理等应用领域中，它是近年来在工具和方法上的重大突破。小波变换是一种可同时在时频两域表征信号局部特征的时频局部化分析方法，即在低频部分具有较高的频率分辨率和较低的时间分辨率，在高频部分具有较高的时间分辨率和较低的频率分辨率，所以被誉为分析信号的“数学显微镜”。由于其具有良好的时间频率分辨率而在许多领域得到广泛应用，因而在金融时间序列等方面必将有十分广阔的应用前景。目前小波分析理论在金融数据分析上的应用大多数只限于小波去噪功能，然而小波真正的强大在于其“多分辨率”功能，即多尺度变换，并且可以同时实现数据去噪与数据约简（降维），本文受此启发，提出了一种基于多尺度小波变换的时间序列相似性度量方法。

1.3 时间序列挖掘研究现状

1.3.1 时间序列挖掘面临的问题

时间序列挖掘已发展成为数据挖掘研究的一个重要分支，受到数据挖掘研究者的广泛关注。从文献综合情况来看，时间序列的挖掘研究目前主要集中在时间序列中相似序列搜索、频繁模式发现、关联模式发现、周期模式发现以及异常数据挖掘等方面[10]。目前时间序列挖掘面临的主要问题有：

(1)时间序列的相似性度量

关于时间序列相似性度量方法人们尽管已经研究得较多，比如直接距离法、特征参数距离法、相关系数法、神经网络学习法、原子序列匹配法等等[1]，但还远没有达到人们所期望的准确度。而相似性度量又是时间序列模式挖掘的基础，如何更好地度量相似性就显得尤为重要，这也是本文致力解决的一个问题。

(2)时间序列的特征提取

对时间序列进行特征提取一方面可以达到数据降维，方便进一步的研究，比如对一个长序列，通过时频变换，就可以用前面几个高能量的系数来刻画，丢失的信息却较少[11]；另一方面，通过特征提取更容易反映序列的本质，比如对序列的幅度、波动频率、趋势、极值点等特征的提取无疑可大大加深对序列的理解和把握，为进一步的分析比较奠定基础。但是时间序列的特征是多方面的，不只是这些常见的特征，它与实际问题是相强关的，更多的特征必须根据问题本身的性质去发现和研究。

(3)时间序列的分割

时间序列的分割是指将一个长时间序列划分为若干个子序列。这也是进行模式挖掘的一个基础性问题，对挖掘的进行与结果有很大影响。但是如何合理的分割，却非易事。目前采取的常见分割方法是等宽度的滑动窗口方法[11]。这种方法的一个明显的缺点是不但效率低，而且由于事实上序列中的模式长度不一，等宽度的一刀切方法显得过于武断，而宽度值的选取也是一个疑问。因此研究自适应的分割方法具有重要的意义。李斌、谭旭都采用了线性化分段的方式实现分割，不失为一种新的尝试和选择[14]。

(4)序列的模式聚类与分类

将序列进行分割，并提取各个子序列的特征后，由于存在多个特征，每个特征的取值可能有多个，每个子序列对应特征空间中的一个点，而挖掘的目的是从复杂的数据中抽取简单的、易被人理解的规律和知识，因而在时间序列挖掘中，经常需要把这些子序列归类成少数几个模式，以便于人们理解和掌握。那么如何进行分类、聚类，这都是有待进一步研究的。

(5)规则的筛选或约简

采用数据挖掘对时间序列进行分析，可能会产生大量的模式或规则。在为数众多的规则中，有相当多的规则显然与应用无关的或者用户已经熟知，只有其中